Final Exam - Moed B
Total time for the exam: 3 hours!

Important: For True / False questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, and provide counterexample otherwise.

1) True or False (24 Points):
   a) True/False: For two random variables, X and Y, \( H(f(X, Y)) \leq H(g(X)) + H(h(Y)) \), where f, g, h are arbitrary functions. (6 pts)
   b) Consider a Gaussian channel where the input, X, has a power constraint P, the noise, \( Z \sim N(0,1) \) and the output is \( Y = X + Z \). The output Y is fed through a function \( f_i(y) = y^i \) where i is an integer. The capacity of this channel is denoted by \( C_i \). Complete \( <, >, = \) between \( C_2 \) and \( C_4 \), prove your answer. (6 pts)
   c) Consider a clean channel with \(|\mathcal{X}| \) inputs and outputs (Fig. 1). Two systems are defined as follows:
      - System A: At each time, a random variable \( Z \sim \text{Unif}[1, \ldots, |\mathcal{X}|] \) determines how many links can be used at the next channel use. This random variable is known to the encoder and the decoder. The capacity of this system is denoted by \( C_A \).
      - System B: In this system, there is a clean channel but with \(|\mathcal{X}'| = \frac{1+\ldots+|\mathcal{X}|}{|\mathcal{X}|} \) inputs (the average amount of inputs) at all times. The capacity of this channel denoted by \( C_B \).
   True/False: The capacity of system B is larger than the capacity of system A, i.e. \( C_B \geq C_A \). (12 pts)

2) Constrained Markov chain (24 Points):
   A random process, \( X_1, X_2, \ldots \) is a Markov chain if it has the Markov property \( X_i = X_{i-1} - X_{i-2} \) for all \( i \geq 3 \). In this question, the Markov chain \( X_1, X_2, \ldots \) takes values from a binary alphabet, \( \mathcal{X} = \{0, 1\} \), and does not contain two consecutive ones (that is, '11' is not valid). The conditional probability, \( P_{X_{i|X_{i-1}}} \), of the Markov chain is given by
   \[
   T = \begin{pmatrix}
   1-p & p \\
   1 & 0
   \end{pmatrix},
   \]
   for all \( i \geq 2 \), where \( p \in [0, 1] \). The matrix rows correspond to \( X_{i-1} \) and the matrix columns correspond to \( X_i \), for example, \( P(X_i = 1|X_{i-1} = 0) = p \). The distribution of \( X_1 \) is to be defined later.
   a) Explain why the Markov chain does not contain consecutive ones.
   b) The stationary distribution of a Markov chain is defined as a probability vector that solves \( vT = v \). Find the stationary distribution of this Markov chain as a function of \( p \).
   - From now on, assume that \( X_1 \) is distributed according to \( v \) that you found in (b).
   c) Compute \( P(X_2 = 0), P(X_3 = 0) \) and \( P(X_7 = 0) \) as a function of \( p \).
   d) (True/False) The entropy rate is defined as \( H(X) = \lim_{n \to \infty} \frac{1}{n} H(X^n) \). Is it true that \( H(X) = H(X_2|X_1) \)?
   e) Compute the entropy rate of the Markov chain as a function of \( p \). (The answer should not contain a limit)
   f) In order to maximize the entropy rate, you can now optimize the parameter \( p \). Does the optimal parameter satisfy \( p = 0.5 \), \( p < 0.5 \) or \( p > 0.5 \)? (You don’t have to solve the maximization problem but you should prove your answer.)
   * Roughly speaking, the amount of sequences of length \( n \) and without '11' is \( 2^n H(X) \). In magnetic storage, such as standard hard disk, it is useful to encode data into constrained sequences (in order to do decrease errors appearances) so the larger the entropy so the better it is.

3) Polarization and the idea of polar codes (28 Points):
   The question is about polarization effect in memoryless channels that can lead to simple coding schemes that achieve the capacity which are called polar codes.
   a) Consider the channel in Fig. 2 where two parallel binary erasure channels can be used at once (the input is \( X = (X_1, X_2) \)). The inputs alphabets are binary, so that \( Y_1 \) and \( Y_2 \) are the outputs of a BEC(p) with inputs \( X_1 \) and \( X_2 \), respectively. Compute the capacity of this channel, namely,
   \[
   \max_{p(x_1, x_2)} I(X_1, X_2; Y_1, Y_2). \tag{1}
   \]
   What is the input distribution \( p(x_1, x_2) \) that achieves the capacity?
   b) Consider the system in Fig. 3, where addition is modulo 2:
Logistic Regression (24 Points)

4) 

where \( f(\hat{\theta}) \) is the main idea of polar codes, which achieves capacity.

Recall the sigmoid function, \( \sigma(x) \).

b) Is the sigmoid function

\[
\sigma(x) = \frac{1}{1 + e^{-x}}
\]

d) Compare the mutual information of Channels 1 and 2 with the capacity of a binary erasure channel (that is, write <, > or \( = \) with simple proof).

*For large \( n \), repeating this decomposition \( n \) times, ends up in \( nc \) clean channels and in \( n(1 - c) \) totally noisy channels. This is the main idea of polar codes, which achieves capacity.

4) Logistic Regression (24 Points):

Recall the sigmoid function, \( \sigma(z) = \frac{1}{1 + e^{-z}} \). The logistic regression classifier, that we’ve learned in class, is a binary classifier.

The estimated probability \( \hat{p}(x^{(i)}; \theta) \) is defined as

\[
\hat{p}(y^{(i)} = 1|x^{(i)}; \theta, b) = h_{\theta,b}(x^{(i)}) = f(\theta^T x^{(i)} + b)
\]  

(2)

where \( f(z) \) is usually the sigmoid function \( \sigma(z) = \frac{1}{1 + e^{-z}} \).

a) Assume that \( \{ (x^{(i)}, y^{(i)}) \}_{i=1}^m \) is a set of i.i.d samples Write the estimated probability of the entire set, i.e., write \( \hat{p}(y^{(i)}, \ldots, y^{(m)}|x^{(1)}, \ldots, x^{(m)}) \) in terms of \( \{ (x^{(i)}, y^{(i)}) \}_{i=1}^m \) and \( h_{\theta,b}(\cdot) \).

b) Is the sigmoid function \( \sigma(z) \) convex, concave or none? Prove your claim.

c) Assume that the sigmoid function is replaced with the following piecewise linear function

\[
\begin{cases}
0 & \text{if } z < -0.5 \\
0.5 + x & \text{if } -0.5 \leq z \leq 0.5 \\
1 & \text{if } z > 0.5
\end{cases}
\]  

(3)

Let \( x = (x_1, x_2) \) be a binary vector, namely, \( x_1 \in \{0, 1\}, x_2 \in \{0, 1\} \). Can you find \( \theta_1, \theta_2 \) and \( b \) such that \( f(\theta^T x + b) \) is the logical or between \( x_1 \) and \( x_2 \)? If yes, do it. If no, prove it doesn’t exist. **Hint:** recall that \( \theta^T x = \theta_1 x_1 + \theta_2 x_2 \).

d) Can you find \( \theta, b \) such that \( f(\theta^T x + b) \) is the logical exclusive or (XOR) between \( x_1, x_2 \)? If yes, do it. If no, prove it doesn’t exist.

Good Luck!